Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

ÉCHANTILLONNAGE ET ESTIMATION AVEC SEUILS D'EXCLUSION

Hanna Elisson[1] et Eva Elvers[2]

RÉSUMÉ

L'échantillonnage avec seuil d'exclusion exclut délibérément une partie de la population cible. Dans le cas de la statistique des entreprises, la base de sondage et l'échantillon se limitent ordinairement aux entreprises d'une certaine taille, c'est-à-dire comptant au moins un certain nombre d'employés. On élimine ainsi le fardeau de réponse des petites entreprises, mais il faut recourir à des hypothèses pour la partie non échantillonnée de la population. Nous présentons ici certains résultats empiriques en fonction d'une enquête et de données administratives suédoises. Nous étudions les différentes sources d'erreur et leur incidence sur l'exactitude des résultats généraux. L'échantillonnage avec seuil d'exclusion est une méthode valable, mais il faut faire preuve de prudence lorsqu'on mesure la taille et procéder à des travaux méthodologiques faisant appel à des modèles.

MOTS-CLÉS : Échantillonnage avec seuil d'exclusion; Données administratives; Estimation par modèle

1. INTRODUCTION

Dans le domaine de la statistique des entreprises, il arrive que les (très) petites entreprises soient exclues de la base de sondage. S'il n'est pas négligeable, l'apport de cette partie de la population est, du moins, faible par rapport à celui de la population restante. Il peut être tentant de ne pas consacrer de ressources à des entreprises qui comptent pour peu dans les résultats globaux de l'enquête. De plus, cette exclusion réduit le fardeau de réponse de ces petites entreprises. Par contre, l'échantillonnage avec seuil d'exclusion peut être considéré comme une méthode inéquitable, simplement parce que la probabilité d'échantillonnage est fixée à zéro. Faut-il considérer l'échantillonnage avec seuil d'exclusion comme une méthode acceptable ou non? Dans l'affirmative, comment faut-il choisir le seuil d'exclusion?

En cherchant à répondre à ces questions concernant l'échantillonnage avec seuil d'exclusion et les seuils pertinents, nous n'avons guère trouvé de sources fondées sur l'expérience et la pratique. Une description de l'échantillonnage avec seuil d'exclusion figure dans l'ouvrage de Särndal, Swensson et Wretman (1992), avec certaines mises en garde. Haan, Opperdoes et Schut (1999) ont une expérience positive de l'échantillonnage avec seuil d'exclusion dans le contexte de l'indice des prix à la consommation. Au Canada, l'Enquête mensuelle sur les industries manufacturières (EMIM) a fait un choix : « La base de sondage de l'EMIM est déterminée à partir de la population cible, après avoir écarté les établissements faisant partie de la tranche inférieure de 2 % de l'estimation totale des livraisons manufacturières pour chaque province. Ces établissements ont été exclus de la base afin de réduire la taille de l'échantillon sans influer significativement sur la qualité. » Cette citation est tirée de la série d'énoncés sur la qualité des données de Statistique Canada (2001).

À Statistics Sweden, plusieurs enquêtes utilisent depuis longtemps l'échantillonnage avec seuil d'exclusion, mais nous n'avons pas trouvé de documentation méthodologique à l'appui. Récemment, on a haussé le seuil d'exclusion de quelques enquêtes afin de pallier les compressions budgétaires. La modification s'est faite très rapidement, voire trop rapidement à notre avis. D'autres enquêtes pourraient connaître le même sort. Le

[1] Hanna Elisson, Statistics Sweden, SCB ES/SES, Box 24 300, SE-104 51 Stockholm, Suède.
[2] Eva Elvers, Statistics Sweden, SCB ES/SES, Box 24 300, SE-104 51 Stockholm, Suède.

sujet de l'échantillonnage avec seuil d'exclusion mérite des études méthodologiques tenant compte des différents avantages et inconvénients concernant la qualité et les coûts. Le recours à l'échantillonnage avec seuil d'exclusion et la fixation du seuil pertinent doivent reposer sur des principes bien définis.

Nous présentons ici les premières constatations de notre étude sur l'échantillonnage avec seuil d'exclusion. Nous avons choisi une enquête suédoise pour illustrer notre propos par des exemples numériques. Comme c'est souvent le cas au cours d'une étude méthodologique, l'étude d'un problème débouche sur d'autres problèmes et d'autres constatations. Dans la section 2, nous décrivons d'abord certaines données suédoises afin de bien situer le contexte de notre étude. Puis, dans la section 3, nous étudions la question de l'exactitude du point de vue de l'échantillonnage avec seuil d'exclusion. La section 4 porte sur les petites et moyennes entreprises et la section 5, sur celles qui comptent peu ou point d'employés. Dans la section 6, nous avons recours à la répartition de Neyman pour évaluer l'importance de différentes tranches de taille et pour comparer les variances. La section 7 présente nos conclusions.

2. LE CONTEXTE SUÉDOIS

2.1 Données administratives, Registre des entreprises et système d'échantillonnage

Le Registre des entreprises (RE) suédois est un registre de plusieurs unités statistiques, notamment l'entreprise, l'unité locale, le genre d'activité et le genre d'activité locale. Une entreprise se compose ordinairement d'une seule entité juridique mais, dans certains cas, elle peut en compter plus d'une. À l'heure actuelle, le RE compte environ 50 entreprises composites totalisant quelque 600 entités juridiques. La pratique en usage en Suède diffère quelque peu de la recommandation de l'Union européenne. Même si les entreprises composites sont peu nombreuses, la plupart sont importantes dans leur branche d'activité. Dans la présente étude de l'échantillonnage avec seuil d'exclusion, nous nous intéressons surtout aux petites et moyennes entreprises; nous nous concentrerons donc sur celles qui comptent une seule entité juridique et un seul genre d'activité.

On trouve des renseignements sur le chiffre d'affaires et l'emploi des entités juridiques en faisant appel aux fichiers administratifs sur la TVA (taxe à la valeur ajoutée) et sur le système de retenue à la source. Une entité juridique est dotée d'un numéro d'identification pour les fins des autorités fiscales et du RE. Toutes les deux semaines, le RE obtient de la Commission nationale de fiscalité des renseignements sur les ajouts et les suppressions d'entités juridiques. Le nombre d'employés est mis à jour à partir de plusieurs sources. Les deux principales sont les renseignements sur le système de retenue à la source, qui servent à calculer le nombre d'employés des entités juridiques à un seul établissement, et un questionnaire du RE envoyé aux entités juridiques à plusieurs établissements. Chacune de ces formalités est remplie essentiellement une fois par année. Statistics Sweden obtient chaque mois des renseignements sur la TVA. L'entité juridique est une entité de base et constitue une source précieuse de renseignements. Toutefois, différentes entités juridiques peuvent déclarer la TVA et la retenue à la source pour une seule et même activité. C'est le cas des entités juridiques faisant partie d'une entreprise composite, mais aussi d'autres groupes d'entités juridiques.

Pour les fins de la statistique des entreprises, Statistics Sweden possède un système d'échantillonnage que la plupart des enquêtes courantes utilisent pour prélever des échantillons. Le système a recours à des nombres aléatoires permanents. Il s'agit d'une méthode de coordination pratique et souple. Il existe une coordination positive et négative entre les enquêtes. Les échantillons successifs de chaque enquête se chevauchent pendant un temps assez long, ce qui est favorable lorsqu'on estime les variations. Toutefois, on procède par rotation afin de réduire le fardeau de réponse.

Bon nombre d'échantillons à utiliser pour les statistiques à court terme de l'année t sont prélevés en novembre de l'année (t–1). Les renseignements contenus dans la base de sondage sont assez récents dans le cas de certaines variables, soit la fin de septembre pour les entreprises et les unités locales actives. Les autres renseignements sont moins récents : le nombre d'employés renvoie au printemps de l'année (t–1) pour les entreprises à plusieurs établissements et à la fin de l'année (t–2) pour les entreprises à un seul

établissement (il s'agit, respectivement, des questionnaires du RE et des renseignements sur la retenue à la source). À partir de 2001, on créera également des bases de sondage en mars. Le principal avantage est la prise en compte d'un nombre considérable de réorganisations au 1er janvier. Dans le RE, les entreprises à un seul établissement ne comptent ordinairement aucun employé pour l'année de l'ajout et ce, jusqu'en mai de l'année suivante. Ce délai contribue à l'hétérogénéité de la tranche de taille des entreprises à zéro employé. Ordinairement, les enquêtes qui exigent, par exemple, un minimum de 10 employés excluent les ajouts de l'année précédente.

Les méthodologistes recommandent fortement que les échantillons soient mis à jour plus d'une fois par année. On a de plus en plus tendance à comprendre et à suivre ce conseil. Certaines enquêtes à court terme prélèvent des échantillons deux fois par année. Comme certaines variables importantes du RE ne sont mises à jour que quelques fois par année, voire une seule fois, il n'est guère avantageux de renouveler l'échantillon très souvent.

2.2 Exemple numérique tiré d'une enquête

L'enquête que nous utilisons comme exemple dans nos calculs a fait l'objet d'un certain nombre de modifications au cours des cinq dernières années. Il y avait auparavant trois enquêtes distinctes : une enquête mensuelle sur les nouvelles commandes et livraisons, une enquête trimestrielle sur les stocks et une enquête trimestrielle sur l'utilisation de la capacité. En 1998, ces trois enquêtes ont été intégrées en une seule, qui comporte un questionnaire mensuel et un questionnaire supplémentaire tous les trois mois. L'intégration visait notamment à mettre l'accent sur la production (par opposition aux livraisons). Le questionnaire mensuel comprend le nombre de jours de production et la valeur de la production, ainsi que les valeurs des livraisons, des nouvelles commandes et du carnet de commandes (réparties entre le marché intérieur et les exportations). Le questionnaire trimestriel « supplémentaire » comprend les valeurs des différents types de stocks à la fin du trimestre et les variations enregistrées au cours de la période. La présentation du questionnaire fait ressortir les relations entre les variables pour aider le répondant et pour réduire les erreurs de mesure.

Les statistiques couvrent l'industrie minière (mines et carrières) et la fabrication. La classification industrielle utilisée par la Suède est fondée sur la classification européenne NACE rev. 1 (Classification statistique des activités économiques dans la Communauté européenne). Le code NACE comporte deux lettres et quatre chiffres et la version suédoise comporte, dans certains cas, un cinquième chiffre. L'unité d'observation correspond à peu près à l'unité fondée sur le genre d'activité (la différence, attribuable à des raisons historiques, est en voie de disparition). Les variables, dont la valeur des livraisons, ont trait à la fabrication propre et non au commerce. Le niveau de détail des statistiques est déterminé par les besoins des utilisateurs, notamment ceux des Comptes nationaux et d'Eurostat. Le plan d'enquête dépend de ces attentes et des possibilités de collecte des données. On utilise l'échantillonnage aléatoire simple stratifié. En tout, 49 branches d'activité sont recoupées avec six tranches de taille en fonction du nombre d'employés. Le tableau 1 montre toutes les tranches de taille, y compris les tranches à seuil d'exclusion.

Tableau 1. Tranches de taille et échantillonnage

Numéro de la tranche de taille	0	1	2	3	4	5	6	7	8
Nombre d'employés	0	1-4	5-9	10-19	20-49	50-99	100-199	200-499	500–
Échantillon	Seuil d'exclusion			À tirage partiel				À tirage complet	

À la fin des années 1990, le taux de non-réponse a augmenté considérablement. Plusieurs enquêtes avaient alors du mal à maintenir le taux de réponse. Dans le cas de l'enquête qui nous intéresse, l'adoption de nouveaux questionnaires a aggravé la situation. La non-réponse partielle est particulièrement élevée en ce qui concerne la variable « production ». Pour de nombreux répondants, les renseignements demandés ne

font pas partie du système comptable ordinaire. On a pris certaines mesures pour réduire le fardeau de réponse, principalement en réduisant le nombre de variables dans le cas des très petites entreprises, qui reçoivent maintenant un questionnaire simplifié. L'enquête fait l'objet de pressions visant à réduire le fardeau de réponse, surtout dans le cas des petites entreprises, ainsi que de pressions visant à réduire la charge de travail des enquêteurs tout en améliorant la qualité. Afin de ménager les ressources, les responsables de l'enquête ont suggéré de hausser le seuil d'exclusion. Dans la tranche de taille de 10 à 19 employés, la non-réponse est élevée tant au niveau des unités qu'à celui des questions.

Nous travaillons avec des données d'enquête et des données administratives. Il n'y a pas de données d'enquête pour les plus petites tranches de taille, et très peu d'observations juste au-dessus du seuil d'exclusion actuel. Une comparaison entre le chiffre d'affaires d'après la TVA et la variable d'enquête « livraisons de produits fabriqués » révèle un certain nombre de faits. Dans la plupart des cas, les chiffres sont très semblables. Le chiffre d'affaires est un peu plus élevé, notamment en ce qui concerne les grandes entreprises et les entreprises composites, comme le laissent prévoir les définitions. Pour une petite proportion d'entreprises, l'écart entre les valeurs est appréciable. Il est probablement attribuable à différentes utilisations des unités déclarantes.

3. EXACTITUDE ET SOURCES D'INEXACTITUDE LIÉES À L'ÉCHANTILLONNAGE AVEC SEUIL D'EXCLUSION

3.1 Introduction

En matière d'exactitude, nous établissons une distinction entre erreurs systématiques et erreurs aléatoires. Nous tâchons d'adopter des méthodes sans biais et présentant une faible variance. L'erreur quadratique moyenne constitue une façon de résumer les deux types d'inexactitude. Pour analyser et décrire l'exactitude, il est pratique de travailler avec les sources d'inexactitude. On en compte six, qui sont liées aux éléments suivants : échantillonnage, couverture de la base de sondage, mesure, non-réponse, traitement des données et modélisation. L'apport de ces sources d'inexactitude dépend de nombreux facteurs, dont la taille des entreprises comprises dans l'échantillon. Quand on parle de l'échantillonnage avec seuil d'exclusion, on peut faire abstraction des grandes entreprises, puisqu'elles sont toujours comprises.

Nous comparons deux grandes possibilités en matière d'enquête. La première est celle d'une enquête par sondage auprès de l'ensemble de la population, qui utilise l'échantillonnage aléatoire stratifié en fonction de la branche d'activité et de la tranche de taille. La deuxième possibilité est aussi celle d'une enquête par sondage, mais limitée à la population au-dessus d'un seuil d'exclusion et complétée par une estimation par modèle pour la partie exclue. L'enquête dispose d'un budget et il s'agit d'obtenir une exactitude aussi grande que possible en fonction de ce budget. Nous nous intéressons à l'incidence de la mise en œuvre de l'échantillonnage avec seuil d'exclusion et à celle d'une hausse du seuil d'exclusion.

3.2 Sources d'inexactitude et dépendance à l'égard de la taille

Si l'on met en œuvre l'échantillonnage avec seuil d'exclusion, la population de la base de sondage diminue; lorsqu'on hausse le seuil d'exclusion, elle baisse encore. Si la taille de l'échantillon est la même, l'inexactitude due à l'***échantillonnage*** diminue aussi. La taille de l'échantillon n'est pas nécessairement identique mais, compte tenu d'un budget donné, la variation de la taille de l'échantillon est habituellement faible.

Les lacunes de la ***couverture de la base de sondage*** – les écarts entre la population de la base de sondage et la population cible – sont largement attribuables aux retards dans la réception des renseignements. Il reste aussi les erreurs proprement dites. La population des entreprises évolue rapidement : ajouts, suppressions, réorganisations, modifications des activités et de la taille. Le surdénombrement et le sous-dénombrement dépendent de la taille, et les proportions peuvent être passablement élevées dans le cas des petites

entreprises. Lorsqu'on met en œuvre l'échantillonnage avec seuil d'exclusion – par exemple un minimum de z employés –, la couverture de la population enquêtée doit être mesurée en fonction de cette restriction. Les entreprises enquêtées qui comptent moins de z employés sont surdénombrées, alors que les entreprises non enquêtées qui comptent au moins z employées sont sous-dénombrées. Dans une étude antérieure, Elvers (1993) a constaté que cette source d'erreur était appréciable dans le cas d'une enquête suédoise sur les placements dont le seuil d'exclusion était fixé à 20 employés. Les valeurs moyennes des strates observées étaient trop faibles dans les plus petites tranches de taille. Ces tranches de taille comptaient pour une proportion modérée des placements, mais l'incidence des lacunes de la couverture était importante. On a donc modifié le calendrier d'établissement des bases de sondage.

Lorsqu'on pratique l'échantillonnage avec seuil d'exclusion, une façon simple et sommaire de procéder consiste à négliger la partie de la population qui est exclue. Sans qu'on le précise nécessairement, il s'agit d'un genre de ***modélisation***. Lorsqu'on estime un total, l'hypothèse implicite est celle d'un « apport négligeable ». Lorsqu'on estime un indice de variation, l'hypothèse implicite est la suivante : « le taux de croissance est le même de part et d'autre du seuil ». Dans les deux cas, une méthode de rechange consiste à construire un modèle explicite à partir des données de l'enquête et de celles du registre, ce qui devrait permettre d'améliorer l'exactitude. On peut estimer l'exactitude à l'aide du modèle lui-même ou de renseignements externes. Les estimations et les évaluations peuvent être améliorées par la suite, lorsqu'on disposera de plus de renseignements.

En ce qui concerne les sources d'inexactitude liées à la ***non-réponse***, à la ***mesure*** et au ***traitement des données***, la dépendance à l'égard de la taille se manifeste de différentes façons. Souvent, le taux de réponse final augmente en fonction de la taille de l'entreprise mais, comme ce phénomène est partiellement attribuable à la stratégie des rappels, le coût dépend aussi de la taille. Il est parfois difficile d'obtenir des réponses pertinentes, par exemple dans le cas d'une faillite, lorsque l'activité n'est que partielle. Encore une fois, il existe une dépendance à l'égard de la taille en ce qui concerne la qualité obtenue et le coût des rappels, des suivis, etc. Nous disposons de peu de faits précis au sujet de ces aspects de la qualité, et encore moins au sujet de la dépendance à l'égard de la taille.

Le raisonnement qui précède montre que la qualité obtenue et le coût dépendent de la taille des entreprises échantillonnées, mais nous n'en savons pas beaucoup plus. Dans notre première étape, nous nous concentrons sur l'erreur d'échantillonnage et sur le modèle dont on a besoin lorsqu'on pratique l'échantillonnage avec seuil d'exclusion. À titre de première approximation, nous envisagerons également la taille globale de l'échantillon à déterminer selon le budget. Nous supposons donc que la taille de l'échantillon est la même, que l'on pratique ou non l'échantillonnage avec seuil d'exclusion. Comme nous l'avons déjà mentionné, en y regardant de plus près, il y aura des écarts selon la nécessité de deuxièmes contacts, de rappels, de vérifications au sujet de suppressions et d'autres causes de surdénombrement, etc. Il semble raisonnable qu'il existe des écarts entre les tranches de taille, dont certains sont assez considérables pour être pris en compte.

4. CHIFFRE D'AFFAIRES PAR EMPLOYÉ DANS LE CAS DES PETITES ENTREPRISES

Comme nous l'avons mentionné plus haut, les responsables de l'enquête envisagent de hausser le seuil d'exclusion en le faisant passer de 10 à 20 employés. On peut se demander s'il est possible de calculer une estimation par modèle du chiffre d'affaires total des entreprises comptant de 10 à 19 employés à l'aide de renseignements sur les entreprises comptant de 20 à 49 employés. Selon une hypothèse pratique, le chiffre d'affaires par employé serait le même au sein de la branche d'activité. Comme on ne recueille qu'une petite quantité de données dans ces tranches de taille; nous avons utilisé surtout des données administratives (chiffre d'affaires d'après la TVA) dans nos calculs. Nous avons construit deux modèles différents pour estimer le chiffre d'affaires total de la tranche de taille 3.

Modèle 1.
On estime le chiffre d'affaires total de la tranche de taille 3 en multipliant le nombre d'employés de cette tranche de taille par une estimation du chiffre d'affaires par employé de la tranche de taille 4. On utilise la régression simple en prenant le chiffre d'affaires annuel comme variable dépendante y et le nombre d'employés comme variable indépendante x. Le modèle ne comporte pas d'ordonnée à l'origine et la variance résiduelle est proportionnelle à x. Le paramètre de régression, b_{jg}, est estimé par groupe, U_{jg}, où j désigne la tranche de taille et g désigne le groupe d'activité économique. Un groupe d'activité économique se compose d'au moins une strate. L'estimateur par modèle du total Y pour la tranche de taille 3 et le groupe g est exprimé comme suit :

$$\hat{Y}_{3g}^{mod\,1} = \hat{b}_{4g}^{\,t} \sum_{k \in U_{3g}} x_k \quad \text{où} \quad \hat{b}_{4g}^{\,t} = \sum_{h \in U_{4g}} \sum_{k \in s_h} \frac{N_h}{n_h} y_k \Big/ \sum_{h \in U_{4g}} \sum_{k \in s_h} \frac{N_h}{n_h} x_k$$

et s_h désigne l'échantillon prélevé dans la strate h, N_h est la taille de la population et n_h est la taille de l'échantillon. Le temps n'est pas montré explicitement, sauf pour l'année t, dont on a besoin dans le modèle 2.

Modèle 2.
Une analyse approfondie des données de trois années successives montre que les entreprises se distinguent les unes des autres. Il existe des écarts non seulement entre les branches d'activité, mais aussi au sein de ces dernières. Pour certaines entreprises, la valeur du « chiffre d'affaires par employé » est passablement élevée, alors que pour d'autres, elle est assez faible. Lorsque cette valeur est élevée (ou faible) une année, elle est souvent élevée (ou faible) les années suivantes. Pour le même ensemble d'entreprises, le ratio $Q=b_{3g}/b_{4g}$ entre les paramètres des tranches de taille 3 et 4 est donc passablement stable dans le temps. Nous pouvons estimer Q en utilisant le chiffre d'affaires d'après la TVA pour une période antérieure. Nous pouvons appliquer ce Q dans la période en cours où nous avons recueilli des données pour la tranche de taille 4, mais pas pour la tranche 3. Nous le faisons dans le cas des entreprises qui sont comprises dans l'échantillon aux deux moments; ces entreprises sont passablement nombreuses en raison de l'échantillonnage avec coordination positive. Par rapport au premier modèle, nous ajoutons un facteur « supplémentaire » q_k, qui s'écarte de 1 et égale Q dans le cas de ces entreprises. Le deuxième estimateur par modèle du total Y pour la tranche de taille 3 et le groupe g est exprimé comme suit :

$$\hat{Y}_{3g}^{mod\,2} = \hat{b}_{4g}^{\,t} \sum_{k \in U_{3g}} x_k q_k \quad \text{où} \quad q_k = \hat{b}_{3g}^{\,t-1} \Big/ \hat{b}_{4g}^{\,t-1} \quad \text{si} \quad k \in U^{t-1} \quad \text{et } q_k = 1 \text{ dans le cas contraire,}$$

et U^{t-1} est la population au moment *(t-1)*.

Nous avons construit ces deux modèles pour estimer le chiffre d'affaires total de la tranche de taille 3. Certains résultats sont présentés dans le tableau 2. Les deux estimations par modèle sont comparées au chiffre d'affaires total connu. On effectue la comparaison à deux niveaux : la tranche de taille 3 et les tranches de taille 3 à 8, soit les entreprises comptant respectivement de 10 à 19 employés et au moins 10 employés. Les cinq branches d'activité ont été choisies pour montrer une variation des résultats. Les employés de la tranche de taille 3 comptent pour plus de 17 % du nombre total d'employés des entreprises des trois branches d'activité portant les codes NACE 18, 19 et 28. La proportion baisse à 1 % dans le cas des branches d'activité NACE 21 et 34 – ces dernières ne sont pas très sensibles au choix d'un modèle. Les branches d'activité NACE 18 et 19 enregistrent des totaux très faibles et, de ce point de vue, sont moins importantes que la branche d'activité NACE 28.

Le modèle 2 est, dans l'ensemble, supérieur au modèle 1. Les écarts sont très faibles dans la branche d'activité NACE 34, non seulement pour « la totalité » de cette dernière, mais aussi pour la tranche de taille 3. Le modèle 2 donne également de bons résultats dans le cas de la branche d'activité NACE 28, qui appartient au groupe qui compte une proportion élevée de petites entreprises. Il convient de préciser que l'importance de l'erreur due à l'échantillonnage est au moins dix fois supérieure à l'écart entre le total connu et l'estimation par modèle.

Tableau 2. Résultats numériques d'après les estimations par modèle

NACE			Entreprises comptant de 10 à 19 employés			Entreprises comptant au moins 10 employés		
			Chiffre d'affaires total	Estimation du Chiffre d'affaires total	Écart en %	Chiffre d'affaires total	Estimation du Chiffre d'affaires total	Écart en %
18	Fabrication de textiles et de produits textiles	Modèle 1	401	390	0,03	1 781	1 770	0,01
		Modèle 2		327	0,18		1 707	0,04
19	Tannage et finissage du cuir	Modèle 1	233	129	0,45	1 167	1 063	0,09
		Modèle 2		136	0,42		1 070	0,08
21	Fabrication de pâte, de papier et de produits du papier	Modèle 1	1 126	1 201	0,06	102 453	102 528	0,0007
		Modèle 2		1 208	0,07		102 536	0,0008
28	Fabrication de produits métalliques ouvrés	Modèle 1	10 059	10 890	0,08	71 057	71 888	0,01
		Modèle 2		10 255	0,02		71 253	0,003
34	Fabrication de véhicules automobiles, de remorques et de semi-remorques	Modèle 1	1 023	940	0,08	193 455	193 372	0,0004
		Modèle 2		1 024	0,001		193 456	0,00001

5. DIFFICULTÉS LIÉES AUX PETITES ENTREPRISES

Lorsque nous construisons un modèle d'estimation, nous recherchons des caractéristiques que nous pouvons utiliser d'une taille à l'autre et qui sont fondées sur les renseignements disponibles dans les registres. Lorsque nous calculons le ratio entre le chiffre d'affaires et le nombre d'employés, nous utilisons des données faciles à obtenir, mais nous sommes conscients du fait que le dénominateur n'est pas pertinent. Nous voulons obtenir une mesure du travail effectué, et le nombre d'employés est un substitut qui présente plusieurs lacunes. Il peut y avoir d'autres travailleurs qui ne sont pas des employés – par exemple, les propriétaires indépendants – et la quantité de travail par personne varie, puisqu'une personne peut travailler à temps plein, à temps partiel, sur une base temporaire, etc. On trouve dans le RE des variables, dont la forme juridique, qui fournissent certains renseignements. Nous devons modifier le nombre d'employés – surtout lorsque ce nombre est zéro – pour obtenir un chiffre plus représentatif du nombre de travailleurs. Nous ajoutons simplement au nombre d'employés un nombre égal à un pour les entreprises sans employés et qui diminue avec le nombre d'employés jusqu'à zéro pour les entreprises comptant dix employés.

Les ensembles de données sur la TVA et sur la retenue à la source renvoient à des périodes différentes. Si l'entité juridique a subi des modifications (fusions, scissions, etc.), le ratio entre le chiffre d'affaires et le nombre (modifié) d'employés n'est pas significatif. L'appariement des numéros d'identification est purement formel, puisque les valeurs du numérateur et du dénominateur ne concordent pas. Certaines de ces discordances formelles sont temporaires, alors que d'autres durent bien des années. Des groupes d'entités juridiques choisissent de déclarer leur chiffre d'affaires aux autorités fiscales là où cela leur convient pour

des raisons fiscales. L'entreprise se prête mieux à l'appariement que l'entité juridique, mais il reste de grandes discordances en ce qui concerne les activités. Nous étions conscients des problèmes de discordance avant d'entreprendre notre étude, mais le nombre d'unités et les valeurs correspondantes sont plus élevés que prévu. Lorsque nous avons exécuté le programme de répartition décrit à la section 6, l'une de nos premières constatations a été l'incidence des valeurs extrêmes. Parmi les entreprises sans employés, on trouve des valeurs du chiffre d'affaires correspondant à 500 et à 50 employés. Bien qu'il n'y en ait pas tellement, il s'agit d'un problème qui a une incidence sur notre étude des seuils d'exclusion ainsi que sur plusieurs autres enquêtes.

Il est nécessaire d'analyser les valeurs extrêmes et les entités juridiques à prendre en compte. Certaines des valeurs élevées du chiffre d'affaires (mais non les valeurs extrêmes) se situent dans la tranche de taille des entreprises sans employés depuis un certain temps. Les valeurs les plus extrêmes s'expliquent par une réorganisation récente. Nous croyons que la plupart des enquêtes ont décelé la plupart de ces variations, mais il peut y avoir des retards. Lorsqu'on a lancé un nouveau RE en 2000, l'un des objectifs était l'harmonisation. Le flux de l'information entre le RE et les enquêtes s'est nettement amélioré, mais on pourrait aller plus loin. Nous en arrivons également à la conclusion que les méthodologistes devraient examiner le plan d'enquête. Peut-être devrait-on utiliser le chiffre d'affaires d'après la TVA. Il s'agirait simplement d'ajouter un groupe spécial composé d'entreprises qui enregistrent un chiffre d'affaires élevé et qui comptent peu d'employés. Dans ce cas, il faut faire attention d'éviter un double compte. Une autre possibilité consisterait à faire du chiffre d'affaires la seule ou la principale mesure de la taille. Avant de modifier l'enquête, il faut procéder à une étude attentive en se demandant où et comment ces variables sont mises à jour dans le RE et quel est leur lien avec les variables d'enquête.

Dans la section 5, nous avons calculé un modèle pour les entreprises comptant de 10 à 19 employés. Avant de construire un modèle des petites entreprises, il faut régler le problème des valeurs extrêmes et élevées. Les petites tranches de taille, notamment celle des entreprises sans employés, sont hétérogènes. Un petit nombre d'entre elles affichent des valeurs élevées pour le chiffre d'affaires, alors qu'un grand nombre enregistrent des valeurs faibles. Le ratio simple entre le chiffre d'affaires et le nombre d'employés modifié pris par tranche de taille augmente pour les tranches de taille au sein de la branche d'activité.

6. ESSAI PAR LA RÉPARTITION DE NEYMAN

Pour les fins de notre exemple, nous avons créé une base de sondage « complète ». Constituée en mai 2001, elle présente les valeurs du chiffre d'affaires d'après la TVA pour l'année 2000. Nous avons éliminé de nos calculs les entreprises sans TVA et sans employés. Dans les autres cas, nous avons remplacé la valeur manquante de la TVA par une valeur nulle du chiffre d'affaires. L'industrie minière (mines et carrières) et la fabrication comptent ainsi quelque 31 000 entreprises. Nous avons 49 branches d'activité que recoupent les neuf tranches de taille ordinaires mentionnées dans la section 2.2. Les entreprises comptant de 0 à 9 employés se situent en deçà du seuil d'exclusion actuel et figurent pour environ 5 % du chiffre d'affaires et pour 75 % des entreprises.

Pour évaluer l'importance des différentes tranches de taille, nous avons simplement utilisé la répartition de Neyman pour une taille donnée de l'échantillon. À cette fin, nous utilisons le chiffre d'affaires d'après la TVA et nous « éliminons » le problème des valeurs extrêmes décrit dans la section 5 en créant une tranche de taille supplémentaire où nous échantillonnons toutes les unités. Il s'agit d'une façon simple de conserver les valeurs du chiffre d'affaires par branche d'activité sans s'inquiéter de leur emplacement exact et pertinent (unité statistique et tranche de taille). Comme nous l'avons mentionné dans la section 5, ce groupe d'entreprises devrait faire l'objet d'une étude. La variable « chiffre d'affaires » est proche de l'une des variables d'enquête. Selon la règle de répartition de Neyman, à chaque strate correspond une taille de l'échantillon qui est proportionnelle au produit de la taille de la population de la strate par l'écart-type de la variable de répartition pour la population de la strate (voir, par exemple, Särndal, Swensson et Wretman, p. 106).

Nous visons une taille de l'échantillon total qui soit semblable à la taille actuelle. Nous ne nous préoccupons pas encore des coûts. Sur le plan de la précision, nous avons utilisé jusqu'ici les branches d'activité qui se situent au niveau à deux chiffres du code NACE. Nous avons formulé une règle et nous avons exécuté le programme de répartition sans seuil d'exclusion et avec quelques seuils d'exclusion différents selon les tranches de taille ordinaires. La taille de l'échantillon total est la même chaque fois, soit 2 200 entreprises. Ce chiffre ne comprend pas les valeurs extrêmes, mais il est inférieur à la taille actuelle de l'échantillon pour assurer une marge de sécurité. Pour établir les comparaisons, nous procédons par étapes.

La première étape consiste à exclure la plus petite tranche de taille, celle des entreprises sans employés. On observe alors une nette diminution de la variance de l'échantillonnage pour la plupart des branches d'activité qui se situent au niveau à deux chiffres. Nous pouvons calculer l'écart de la variance entre les deux valeurs et en prendre la racine carrée. Le résultat ainsi obtenu pour chaque branche d'activité peut être interprété comme la « marge » dont nous disposons pour calculer un estimateur par modèle, soit la racine carrée de l'erreur quadratique moyenne pour cet estimateur. Lorsqu'on pratique l'échantillonnage stratifié, on a besoin d'un minimum d'observations dans chaque strate. Dans notre exemple, nous obtenons plus de 400 observations « supplémentaires » pour la population entière qui présente déjà au moins trois observations par strate, et une centaine de ces observations supplémentaires se trouvent dans la tranche de taille 0. Lorsque nous supprimons cette tranche de taille, le nombre d'observations « supplémentaires » diminue de 120. Les intervalles de confiance correspondent à environ 90 % des intervalles antérieurs. Dans la plupart des branches d'activité, la marge de calcul de l'estimateur par modèle est considérable; dans bien des cas, elle est de plusieurs fois 100 %. Toutefois, la marge est plus faible pour certaines branches d'activité et, naturellement, pour l'ensemble de la population. Les chiffres présentés ici sont quelque peu optimistes, puisque nous avons éliminé les cas extrêmes.

À bien des égards, l'incidence de la deuxième étape – qui consiste à exclure également la tranche de taille suivante, celle des entreprises comptant de 1 à 4 employés – est semblable à celle de la première. La variance de l'échantillonnage diminue considérablement et les intervalles de confiance correspondent à environ 80 % des intervalles antérieurs. La marge de calcul de l'estimateur par modèle est plus grande qu'auparavant en valeur, mais beaucoup plus faible en pourcentage. Dans la plupart des branches d'activité, cette faiblesse n'est pas inquiétante, mais pour quelques-unes, il convient d'effectuer une autre vérification.

Dans la troisième étape – qui nous amène au seuil d'exclusion actuel – les constatations sont semblables à celles qui concernent la deuxième étape, mais plus prononcées. Les écarts entre les branches d'activité sont considérables; le seuil de dix employés est élevé pour certaines, mais faible pour d'autres. La nécessité de construire un modèle exact varie en conséquence. Dans certaines branches d'activité qui se situent au niveau à deux chiffres, plus de 10 % du chiffre d'affaires se retrouvent en deçà du seuil. Par contre, la plupart de ces branches d'activité comptent pour une faible proportion du total global. Pour d'autres branches d'activité, la proportion qui se situe en deçà du seuil est minime, soit 1 % ou moins. Ici, nous pouvons comparer les résultats à ceux de la section 4. Dans les trois branches d'activité portant les codes NACE 18, 19 et 28, des proportions élevées du chiffre d'affaires se retrouvent également en deçà du seuil d'exclusion. Les deux branches d'activité NACE 21 et 34 figurent parmi celles qui comptent une faible proportion de petites entreprises et une forte « marge » de calcul de l'estimation par modèle. Pour ces deux branches d'activité, on obtient le même résultat si on hausse encore le seuil en le portant à 20 employés. À cette étape, toutefois, la marge est faible pour la plupart des branches d'activité.

7. CONCLUSIONS

Forts d'une certaine expérience, nous trouvons que l'échantillonnage avec seuil d'exclusion constitue une méthode utile, mais qui nécessite une préparation attentive avant d'être utilisée. Il importe d'obtenir une mesure pertinente de la taille. À cet égard, nous avons trouvé la variable « nombre d'employés » trop faible. Les entreprises sans employés, surtout, constituent un groupe hétérogène; il faut donc envisager un autre

regroupement et étudier d'autres mesures de la taille. Qu'on utilise ou non l'échantillonnage avec seuil d'exclusion, le plan d'enquête doit tenir compte des constatations concernant la taille.

Lorsqu'on fixe un seuil d'exclusion, on doit l'adapter à l'enquête et à son objet. Il ne convient pas de fixer un seul et même seuil pour toutes les branches d'activité, car nous avons observé des écarts importants entre les branches d'activité qui se situent au niveau à deux chiffres du code NACE. Pour les fins de notre étude de cas, on peut hausser le seuil actuel pour quelques branches d'activité soigneusement choisies, mais on doit le conserver, voire le réduire, pour d'autres branches d'activité.

Il n'est pas évident de trouver un modèle pertinent pour les estimations de la partie de la population qui est exclue. Il existe des écarts entre les tranches de taille. Nous l'avons constaté aussi bien dans le cas des très petites entreprises que dans celui des petites et moyennes entreprises. Toutefois, certaines relations entre les variables semblent être (passablement) stables dans le temps, et ces relations peuvent améliorer l'estimateur par modèle. Le modèle utilise alors les renseignements récents tirés de l'échantillon et les relations antérieures tirées des données administratives. Il faut déployer certains efforts pour construire ces modèles, qui peuvent avoir des composantes propres à la branche d'activité et à la taille, et il faut exercer une supervision permanente. Le coût de ces efforts entre dans le rapport qu'il faut établir entre le total des coûts et l'exactitude.

BIBLIOGRAPHIE

Elvers, E. (1993), « A New Swedish Business Register Covering a Calendar Year and Examples of its Use for Estimation », *Proceedings of the International Conference on Establishment Surveys, American Statistical Association,* p. 916 à 919.

Haan, J. De, E. Opperdoes et C.M. Schut (1999), « Le choix des produits pour l'indice des prix à la consommation : le seuil d'exclusion par opposition au sondage probabiliste», Techniques d'enquête, n° 25, p. 33-45.

Särndal, C.-E., B. Swensson et J. Wretman (1992), *Model Assisted Survey Sampling,* New York, Springer-Verlag.

Statistique Canada (2001), « Enquête mensuelle sur les branches d'activité manufacturières », *Système de documentation des données statistiques,* n° de référence 2101, Statistique Canada.